6 Optimal Transport

6.1 Motivation: Bootstrap Consistency

Let $X_1, \dots, X_n \sim F$ be independent random variables taking values in $\mathbb{R}^d$. Let $\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ be an estimator of $\mu = \mu(F) = \mathbb{E}X_1$. We want a confidence region for μ based on μ_n. By the central limit theorem, $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} N_d(0, \Sigma(F))$, where $\Sigma(F) = \mathbb{E}X_1X_1^t$ is the covariance matrix associated to F.

The idea of bootstraping is as follows. We know everything about the empirical distribution function F_n (see section 1). Let $X_1^*, \dots, X_n^* \sim F_n$ be independent (i.e., samples with replacement from the original data). We know that:

$$\sqrt{n}(\mu_n - \mu) \xrightarrow{D} N_d(0, \Sigma(F)), \qquad n \to \infty,$$
$$\sqrt{m}(\mu_m^* - \mu_n) \xrightarrow{D} N_d(0, \Sigma(F_n)), \qquad m \to \infty,$$

where $\mu_n^* = \frac{1}{m}\sum_{i=1}^m X_i^* = \mu(F_n)$ is the **bootstrap sample mean**. Then conditionally on $X_1, \dots, X_n, \dots$, we also know that

$$\Sigma(F_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_n)(X_i - \mu_n)^T \to \Sigma(F)$$

almost surely by the law of large numbers. We want to conclude that $\sqrt{m}(\mu_m^* - \mu_n)$ and $\sqrt{n}(\mu_n - \mu)$ have approximately the same distribution for m, n large, conditionally on $X_1, \dots$. For this we need some uniformity in the convergence. Wasserstein distances are metrics between probability distributions that allow to prove such uniformity in the convergence: we shall see that the distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ is close to the distribution of $\sqrt{m}(\mu_m - \mu)$ where the uniformly in m.

6.2 Wasserstein Distances

For P, Q distributions on $\mathbb{R}^d$ define the **Wasserstein distance** of order $p \geq 1$ by

$$W_p(P, Q) = \inf_{X \sim P, Y \sim Q} (\mathbb{E}(\|X - Y\|^p))^{1/p},$$

where the infimum is taken over all **couplings**, i.e., all joint distributions $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^d$ such that marginally $X \sim P$ and $Y \sim Q$. This is a convex non-empty set, since we can always take X and Y independent. The infimum is always attained (proved on example sheet 3), but here we will not need this result.

For random variables X and Y, let P_X and P_Y denote their distributions. The Wasserstein distance $W_p(P_X, P_Y)$ does not depend on the dependence between X and Y, only on their marginal laws P_X and P_Y.

Example 6.1. *If $X \sim N(0, 1)$ then $W_p(P_X, P_{-X}) = 0$. Even though the realisations of X are different from the realisations of $-X$, the measures are equal: $X \overset{D}{=} -X$. Thus, in*

the optimisation problem defining the Wasserstein distance we can take $Y = X$, so that $Y \sim P_{-X}$ and $0 \leq W_p^p(P_X, P_{-X}) \leq \mathbb{E}(\|X - Y\|^p) = 0$.

Lemma 6.2. *We have the following properties:*

1. *If X and Y are independent then $W_p(P_{X+Y}, P_X) \leq W_p(P_Y, P_0) = (\mathbb{E}(\|Y\|^p))^{1/p}$.*
2. *If $a \in \mathbb{R}$, then $W_p(P_{aX}, P_{aY}) = |a| W_p(P_X, P_Y)$.*
3. *If $(U_1, \ldots, U_n)$ are independent and $(V_1, \ldots, V_n)$ are independent, then*
$$W_p(P_{\sum U_j}, P_{\sum V_j}) \leq \sum W_p(P_{U_j}, P_{V_j}).$$
4. *If in addition (to the conditions in part 3) $\mathbb{E}U_j = \mathbb{E}V_j$ for all j, then*
$$W_2^2(P_{\sum U_j}, P_{\sum V_j}) \leq \sum W_2^2(P_{U_j}, P_{V_j}).$$

Proof. (1). Let X and Y be independent. Then the coupling $(X + Y, X)$ yields

$$W_p(P_{X+Y}, P_X) \leq (\mathbb{E}\|X + Y - X\|^p)^{1/p} = (\mathbb{E}\|Y\|^p)^{1/p}.$$

(2) is proved on example sheet 3.

(3). Let U_j and V_j be such that $(\mathbb{E}\|U_j - V_j\|^p)^{1/p} \leq W_p(U_j, V_j) + \epsilon 2^{-j}$ for all j, such that the pairs $\{(U_j, V_j)\}$ are independent. Then set $U = \sum U_j$ and $V = \sum V_j$, which have the correct marginal distributions. Then by Minkowski's inequality

$$W_p(P_U, P_V) \leq (\mathbb{E}\|\sum U_j - V_j\|^p)^{1/p} \leq \sum (\mathbb{E}\|U_j - V_j\|^p)^{1/p} \leq \epsilon + \sum W_p(P_{U_j}, P_{V_j})$$

taking the limit as $\epsilon \to 0$ gives the result.

(4). Using the same construction as in (3), we have:

$$W_2^2(P_U, P_V) \leq \mathbb{E}\|\sum U_j - V_j\|^2 = \sum \mathbb{E}\|U_j - V_j\|^2 \leq O(\epsilon) + \sum W_2^2(P_{U_j}, P_{V_j})$$

Again taking the $\epsilon \to 0$ limit gives the result. The middle equality comes from $\mathbb{E}(U_j - V_j) = 0$, so the cross terms vanish. □

We call a function $d : X \times X \to [0, \infty]$ a **pseudometric** if it is symmetric, $d(x, x) = 0$ for all $x \in X$ and d satisfies the triangle inequality. This is the same as a metric, without the notion of distinguishability; $d(x, y)$ may vanish for $x \neq y$.

Proposition 6.3. *W_p is a pseudometric, and it is finite if P, Q have finite moments of order p (i.e., $\mathbb{E}(\|X\|^p) + \mathbb{E}(\|Y\|^p) < \infty$ for $X \sim P$ and $Y \sim Q$).*

We shall see that W_p is a metric.

Proof. Finiteness follows from the inequality $(a-b)^p \le 2^p(a^p+b^p)$ for $a, b \ge 0$: For any coupling (X, Y), $\mathbb{E}(\|X-Y\|^p) \le 2^p\mathbb{E}\|X\|^p + 2^p\mathbb{E}\|Y\|^p < \infty$. Symmetry is clear from the definition. Clearly $W_p(P_X, P_X) = 0$. Now for the triangle inequality, let P, Q, R be three distributions and let X_1, Y_1, Y_2, Z_2 be such that $X_1 \sim P$, $Y_1, Y_2 \sim Q$ and $Z_2 \sim R$, and $W_p^p(P,Q) = \mathbb{E}(\|X_1 - Y_1\|^p)$ and $W_p^p(Q,R) = \mathbb{E}(\|Y_2 - Z_2\|^p)$.[1] The gluing lemma (see the appendix and example sheet 3) asserts the existence of (X, Y, Z) such that $(X, Y) \overset{D}{=} (X_1, Y_1)$ and $(Y, Z) \overset{D}{=} (Y_2, Z_2)$. Then

$$W_p(P,Q) \le (\mathbb{E}\|X-Z\|^P)^{1/p} \le (\mathbb{E}\|X-Y\|^p)^{1/p} + (\mathbb{E}\|Y-Z\|^p)^{1/p} = W_p(P,Q) + W_p(Q,R)$$

so W_p is indeed a pseudometric. □

Proposition 6.4. *Assuming that $\mathbb{E}\|X\|^p < \infty$, $W_p(P_{X_n}, P_X) \to 0$ if and only if both $X_n \xrightarrow{D} X$ and $\mathbb{E}\|X_n\|^p \to \mathbb{E}\|X\|^p$.*

Proof. Suppose that $W_p(P_{X_n}, P_X) \to 0$. Let $Y_n \sim P_X$ be such that $W_p^p(P_{X_n}, P_X) = \mathbb{E}\|X_n - Y_n\|^p$. Let f be 1-Lipschitz, then $|\mathbb{E}f(X)| \le \mathbb{E}\|X\| \le (\mathbb{E}\|X\|^p)^{1/p} < \infty$ and

$$|\mathbb{E}f(X_n) - \mathbb{E}f(X)| = |\mathbb{E}(f(X_n) - f(Y_n))| \le \mathbb{E}\|X_n - Y_n\| \le \mathbb{E}(\|X_n - Y_n\|^p)^{1/p} = W_p(P_{X_n}, P_X) \to 0.$$

By the Portmanteau lemma (in the appendix) this implies that $X_n \xrightarrow{D} X$. We show on example sheet 3 that $W_p^p(P_{X_n}, P_0) = \mathbb{E}\|X_n\|^p$, so

$$|(\mathbb{E}\|X_n\|^p)^{1/p} - (\mathbb{E}\|X\|^p)^{1/p}| = |W_p(P_{X_n}, P_0) - W_p(P_X, P_0)| \le W_p(P_{X_n}, P_X) \to 0.$$

We will now show the converse in several steps.

Step 1 — Restriction to a compact set. For all $\delta > 0$ there exists an $R_\delta \in [1, \infty)$ such that for $X_n \sim P_{X_n}$ and $Z_n = X_n 1(\|X_n\| \le R)$, we have $W_p^p(P_{X_n}, P_{Z_n}) \le (1 + 2^p)\delta$ for all n and all $R \ge R_\delta$ (see example sheet 3, question 12). Thus we may assume that X_n and X are supported on $C = \{x \in \mathbb{R}^d : \|x\| \le R\}$ for some finite R (note that this required the uniformity of the bound above).

Step 2 — Discretisation. Now let $\epsilon > 0$. There exists finitely many disjoint nonempty sets $B_1, \dots, B_{N_\epsilon} \subseteq C$ such that $\operatorname{diam}(B_i) = \sup_{y,z \in B_i} \|y - z\| \le \epsilon$ for all i and $\mathbb{P}(X \in \partial B_i) = 0$ for all i.[2] Let $y_i \in B_i$ and define

$$X^\epsilon = \sum_{i=1}^{N_\epsilon} y_i 1(X \in B_i) \quad \text{and} \quad X_n^\epsilon = \sum_{i=1}^{N_\epsilon} y_i 1(X_n \in B_i).$$

[1] We can add ϵ if we do not wish to use that the infimum defining the Wasserstein distance is attained.

[2] To achieve this, first cover C by open balls B_i' of diameter in $[\epsilon/2, \epsilon]$, where the diameter is chosen such that $\mathbb{P}(X \in \partial B_i') = 0$. This is possible, since only for countably many radii the boundary will have positive probability. Then let $B_j = B_j' \setminus \cup_{i<j} B_i'$ and remove the empty B_j's from the resulting collection.

Then

$$W_p^p(P_{X_n^\epsilon}, P_{X_n}) \leq \sum_{i=1}^{N_\epsilon} \mathbb{E}(\|X_n^\epsilon - X_n\|^p 1(X_n \in B_i)) \leq \sum_{i=1}^{N_\epsilon} \epsilon^p \mathbb{E}(1(X_n \in B_i)) = \epsilon^p.$$

Similarly $W_p^p(P_{X^\epsilon}, P_X) \leq \epsilon^p$. If suffices to bound the $W_p(P_{X_n^\epsilon}, P_{X^\epsilon})$. Let $p_i = \mathbb{P}(X^\epsilon \in B_i)$, $q_i = \mathbb{P}(X_n^\epsilon \in B_i)$ and define π by

$$\pi(\{(y_i, y_j)\}) = \begin{cases} q_i & \text{if } i = j \in I \\ p_i & \text{if } i = j \in J \\ \alpha_i \beta_j & \text{if } i \in I, j \in J \\ 0 & \text{otherwise.} \end{cases}$$

where $I = \{i : p_i \geq q_i\} = J^c$, $\alpha_i = p_i - q_i$ and $\beta_j = \frac{q_j - p_j}{\sum_{k \in J}(q_k - p_k)}$. Let $(U, V) \sim \pi$. Then it is a simple exercise to show that $U \sim P_{X^\epsilon}$ and $V \sim P_{X_n^\epsilon}$, so π represents a valid coupling for P_{X^ϵ} and $P_{X_n^\epsilon}$ and therefore

$$W_p^p(P_{X_n^\epsilon}, P_{X^\epsilon}) \leq \mathbb{E}\|U - V\|^p \leq \text{diam}(C)^p \mathbb{P}(U \neq V).$$

Now see that as $X_n \xrightarrow{D} X$ and $\mathbb{P}(X \in \partial B_i) = 0$ for all i we have

$$\mathbb{P}(U \neq V) = \frac{1}{2} \sum_{i=1}^{N^\epsilon} |p_i - q_i| = \frac{1}{2} \sum_{i=1}^{N^\epsilon} |\mathbb{P}(X^\epsilon \in B_i) - \mathbb{P}(X_n^\epsilon \in B_i)| \to 0.$$

Thus

$$\limsup_{n \to \infty} W_p(P_{X_n}, P_X) \leq \limsup W_p(P_{X_n}, P_{X_n^\epsilon}) + W_p(P_X, P_{X^\epsilon}) + W_p(X^\epsilon, X_n^\epsilon) \leq 2\epsilon.$$

for all $\epsilon > 0$. Hence $W_p(P_{X_n}, P_X) \to 0$. □

Corollary 6.5. *If $W_p(P_X, P_Y) = 0$ then $P_X = P_Y$, so W_p is indeed a metric.*

6.3 Application to the bootstrap example

See question 13 in example sheet 3.